·论著·

基于特征分类器集成的糖尿病分类方法

刘丽桑[1]，李强[2]，杨帆[2]，郑哲洲[3]，林雪娟[3]，吴青海[3]

([1]福建工程学院信息科学与工程学院，福州 350118；[2]厦门大学自动化系，厦门 361005；[3]福建中医药大学中医证研究基地，福州 350122)

摘要：目的：结合医用电子鼻技术，探讨糖尿病患者及其口腔呼气的气味图谱特征。方法：选择180例糖尿病患者和100例健康者，用医用电子鼻采集280例口腔呼气的气味图谱，采用基于数据特征划分的方法，用支持向量机和随机森林集成模型对糖尿病患者进行分类预测。结果：①线性核函数的支持向量机（SVM1）分类结果不是很理想，低于多项式核（SVM2）、径向基函数核（SVM3）和随机森林（RF）3种分类器，说明分类超平面显然是非线性的；②集成分类器对糖尿病患者和健康者的气味图谱特征的识别准确率可达88.04%。结论：基于特征划分的分类器集成方法预测性能明显好于单一分类器，为使用医用电子鼻进行糖尿病诊断分析提供了一种有效手段。

关键词：糖尿病；特征分类；模式识别；支持向量机集成

基金资助：国家自然科学基金项目（No.81373552），福建省教育厅A类项目（No.JA14212），福建工程学院科研启动项目（No.GY-Z12079）

Classification method of diabetes based on integration of characteristic classifier

LIU Li-sang[1], LI Qiang[2], YANG Fan[2], ZHENG Zhe-zhou[3], LIN Xue-juan[3], WU Qing-hai[3]

([1]School of Information Science and Engineering, Fujian University of Technology, Fuzhou 350118, China; [2]Department of Automation, Xiamen University, Xiamen 361005, China; [3]Research Base of TCM Syndrome, Fujian University of Traditional Chinese Medicine, Fuzhou 350122, China)

Abstract: Objective: To discuss the profile features of oral odor of diabetic patients based on medical electronic nose technology. Methods: 180 patients of diabetes and 100 healthy people were selected, and the profile features of oral odor of 280 volunteers were collected by using medical electronic nose. The classification forecasting was carried out on diabetic patients by using support vector machine (SVM) and random forest integration model based on partitioning method of data characteristics. Results: ①The classification result of SVM1 was not very good, which was lower than that of SVM2, SVM3 and RF, and the result showed that the classification hyperplane is nonlinear. ②The accurate rate of recognition of integrated classifier on diabetic patients and healthy people is 88.04%. Conclusion: The forecasting performance of classifier integration method based on feature division is superior to that of single classifier significantly, which provided an effective means for the diagnostic analysis of diabetes based on medical electronic nose.

Key words: Diabetes; Feature classification; Pattern recognition; Integration of support vector machine

Funding: National Natural Science Foundation of China (No.81373552), Class A Project of Education Department of Fujian Province (No. JA14212), Scientific Research Foundation of Fujian University of Technology (No.GY-Z12079)

中医药现代化是国家中长期科技发展规划中具有战略意义的研究课题。我国在智能中医诊断信息处理技术方面的研究开始于20世纪70年代中期，主要是开展简单的中医专家系统的开发研究[1]。随着中医理论形式化的深入研究，中医专家系统更多采用突破传统“专家系统”概念的先进人工智能技术。电子鼻的出现结合了先进的智能信息处理技术，建立起一个口腔气味图谱与糖尿病病证之间的数学模型和模式识别体系，体现中医辨证施治的重要特点[2]，尽管从人工智能应用角度上没有根本解决中医内在辨证论治原理的描述问题，但是通过电子鼻技术可以实现闻诊在线控制和辅助医生诊断，为人工智能新技术在临床疾病诊断中的研究与应用提供新平台和开拓新思路。

通讯作者：林雪娟，福建省福州市闽侯上街大学城邱阳路1号福建中医药大学中医证研究基地，邮编：350122，电话：0591-2286153

E-mail：lxjfjzy@126.com

近年来，机器学习在临床诊断的作用越来越广泛，使用的分类器有支持向量机、决策树、人工神经网络等[3-5]。集成分类器因其具有很强的泛化能力，能够提高分类精度和稳定性，已成为机器学习的热点[6-8]。本研究采用电子鼻临床检查数据与特征分类器集成相结合的方法建立计算机辅助糖尿病诊断模型，取得了满意的效果。

相关理论

1. 中医电子鼻　电子鼻作为一种非侵入性的医学检测工具，在肺部疾病、糖尿病、尿毒症检测和细菌感染等方面具有重要的临床应用价值。电子鼻的工作原理是模拟人的嗅觉对被测气体进行感知、分析和识别。将电子鼻技术应用于糖尿病的诊断，具有早期、无创、便捷的特点，因此备受医生及患者的青睐。糖尿病的并发症之一是会导致酮酸中毒，它与血液中酮酸体的含量有直接关系。而酮酸体在血液中代谢的最终产物丙酮可以通过呼吸排出体外，通过这一关系利用电子鼻技术直接检测呼气中的丙酮含量就可以间接了解血糖情况，以便时刻监测血糖的变化。1995年王平等[9]利用电子鼻对从医院采集的糖尿病患者的呼出气体进行识别，得到血糖值和呼出气体中丙酮含量呈线性相关性，从而论证了应用呼气诊断糖尿病的可行性。

本研究采用的电子鼻（WES-ENO11103-A）是由吴青海教授自主研发的，包括控制主机、采样装置和计算机软件与显示3个部分[10]，其工作过程可归纳为：传感器阵列→预处理电路→神经网络和各种算法→计算机识别[11]。当气体通过测量气室时，10个传感器同时对同一被测气体进行探测，产生10条不同颜色的响应曲线随时间输出，形成该被测气体的特征气味图谱（见图1）。每条曲线振幅、上升快慢都是电子鼻对某种被测气味响应特性的体现。本研究使用的数据集是对280个研究对象（180例糖尿病患者，100名健康者）按照一定时间顺序采集400次得到的时间序列。其中，病例来源于2013年1月–2013年8月间福建中医药大学附属人民医院、福建省福州中西医结合医院的住院患者。健康者来源于福建中医药大学附属第二人民医院体检中心的健康志愿者。

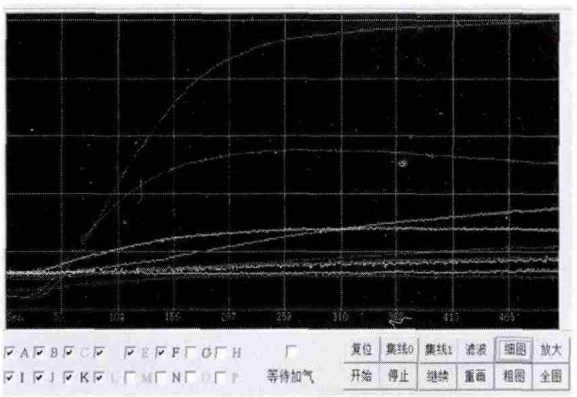

图1　口腔呼气图谱

2. 随机森林　分类回归树（classification And regression tree, CART）是一种以吉尼指数作为内部节点分类标准的算法，可用于离散类别变量的数据分析。在建立CART树时，每个分类属性的选择是根据它在不同预测下对样本数据划分的好坏程度来进行的。随机森林（random forest, RF）是一种以CART为元分类器的组合分类器算法[12-14]。它采用Bagging方法制造有差异的训练样本，并且在构造单棵树时随机地选择特征对内部节点进行属性分类，采用简单多数投票法得到最终输出。Bagging方法和CART算法的结合，使得随机森林能够较好容忍噪声，从而具有较好的分类性能，所以本研究使用随机森林作为一个子分类器。

3. 支持向量机　支持向量机（support vector machine, SVM）是Vapnik等人1995年首先提出的一种以统计学理论为基础的通用学习方法，能够很好地解决小样本、非线性及高维模式识别等实际问题。用于分类问题其实就是寻找一个最优分类超平面作为分类决策面，在高维空间中对类别进行分割，以保证最小的分类错误率，即结构风险最小[15]。

从图1可以看到，图谱中含有400个时间点，每个时间点有10个传感器输出，构成4 000维的高维数据；且有几条曲线振幅小、上升较慢、重叠区域大，线性不可分；因此支持向量机能很好的解决该组数据的识别问题。本研究在线性可分问题描述基础上，增加1个松弛项$\zeta_i \geqq 0$，使得$\min \Phi(w, \xi) = \frac{1}{2}\|w\|^2 + C\left[\sum_{j=1}^{n} \xi_j\right]$，也就是构造一个最优超平面[5-6]。

完整的支持向量机还包括通过核函数的非线性变换将输入空间变换到一个高维空间，然后再高维空间中求取线性分类面。

目前，常用的核函数有3类：线性核函数、多项式核函数、径向基核函数（radial-basis function, RBF）。

分类器集成

集成学习的主要思想是将多个分类器的单独决策以某种方式结合起来，通常能够得到比单个分类器更强的泛化能力。本研究提出的基于特征集成分类器糖尿病识别系统的结构如图2所示。系统的基本思想是充分考虑电子鼻的数据特性，按照时间顺序

和随机两种方式将特征非交叠划分为K个各不相同的特征子集SUB_TR，目的就是为了构造差异性大的集成成员分类器，然后对K个特征子集分别用决策树和支持向量机两种分类器进行初步分类，再用基于特征分类器模型对K个特征子分类器SUB_C的输出采用投票表决法进行融合，实现糖尿病的最终分类，从而提高识别率。

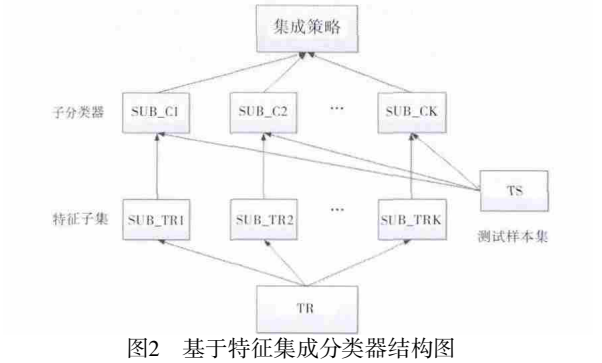

图2　基于特征集成分类器结构图

上述特征集成分类器糖尿病识别系统采用非交叠特征划分方法，用不同的特征训练不同的子分类器，然后将不同的子分类结果融合起来，尽可能加大子分类器之间的差异性，有利于分类器互补，能够保证分类的准确性和稳定性。

结果与分析

本研究硬件环境是：联想服务器Intel™ 3.4G CPU，4G RAM；软件环境是：Window XP系统；所用工具软件：MATLAB 2010b，Libsvm和VC++6.0。

实验数据选择医用电子鼻按时间顺序采集400次得到的数据集，该数据提供了180例糖尿病患者样本和100名正常人样本，每个样本都包含400个序列共4 000维，y为模式类别，其中，$y(x)=0$表示正常人，$y(x)=1$表示糖尿病患者。

首先对训练数据进行归一化(记为S)。由于医用电子鼻每次都测量气体或气味10种成分，按照时间顺序共测量400次，气体或气味的每种成分测量值都构成一个时间序列，然后得到10个时间序列中每个序列的统计特征(均值，均方差)。但均值和方差都是时域特征，并非都能很好的用来分类预测。而小波分解后的低频分解系数能够表征数据的特征信息，小波包分解则是根据需要对小波分解过程中没有分解的高频信号进行再分解，选择合适的基波函数，然后对在这样的最佳小波包基函数下分解后的信号进行时频分析[16]。所以利用小波分析方法对每个序列进行小波分解提取低频系数(记为W1)和小波包分解得到的分解系数(记为W2)。将归一化的数据和小波方法提取的数据分别用于单一分类器，与简单集成分类器进行对比。

本研究提出的集成方法共采用了两种数据特征划分的方式，第1种考虑到医用电子鼻采集到的数据特性，将每次采集的10维序列值作为子训练集(记为EN1)，第2种是将4 000维数据随机划分为400个子训练集(记为EN2)。这两种方式构造的子训练集分别作为子分类器的输入向量。

为了让数据得到充分的利用，SVM模型和随机森林模型(记为RF)均采用的是十折交叉的方法(即将输入向量分成平均10份，轮流将其中9份作为训练数据，1份作为测试数据，进行试验，每次试验都会得出相应的正确率，10次的结果的正确率的平均值作为最后的正确率)。采用SVM模型时需要选择核函数，目前还没有很好的理论依据说明如何选择合适的核函数，主要还是依靠经验，所以本课题组对上述3种核函数都进行了试验(线性核记为SVM1，多项式核记为SVM2，RBF核记为SVM3)。另外试验中采用K折交叉验证(K-CV)来选择惩罚参数C以及RBF的参数γ的最佳取值[17]，K取值为10。实验结果如图3–图6所示。

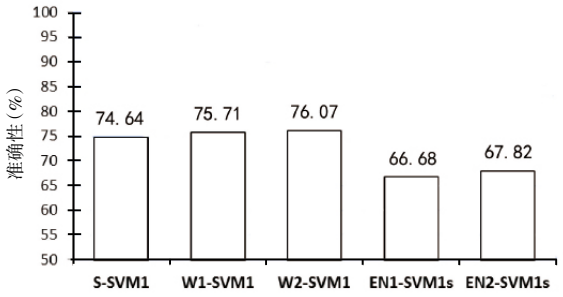

图3　SVM1集成的预测结果比较

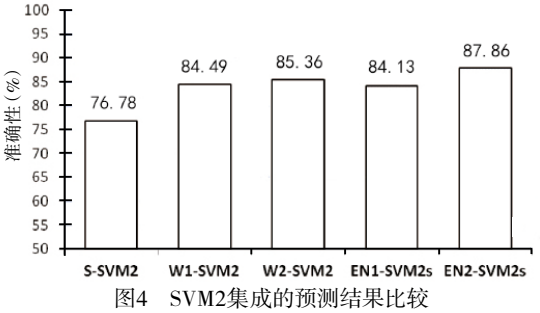

图4　SVM2集成的预测结果比较

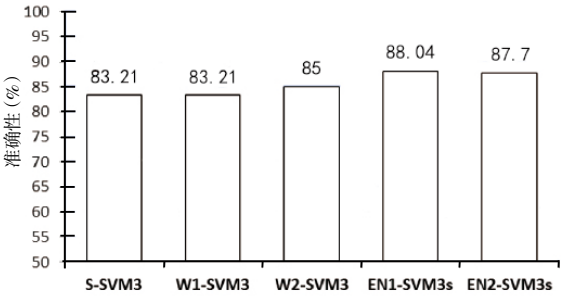

图5　SVM3集成的预测结果比较

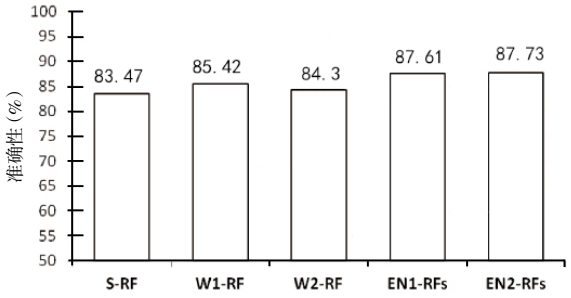

图6 RF集成的预测结果比较

本研究采用分类准确率作为测评指标，该值越高则分类效果越好。实验结果表明，相比SVM2、SVM3和RF 3种分类器，SVM1的分类效果都不是很理想，可以说明分类超平面显然是非线性的。对于SVM2、SVM3和RF 3种分类方法的总体分类效果尚可，同时本研究提出的两种特征划分构造的集成分类器准确率达到87.86%、88.04%和87.73%，分类效果明显好于其他方法。总体来看，使用多项式核和RBF核对于不同的输入向量，本课题组提出的基于特征划分的集成方法是一种简单且效果较好的选择。

小结

本文提出了一种基于特征分类器集成的糖尿病分类方法。针对医用电子鼻应用的特殊性，非线性分类超平面采用多项式核和RBF核函数是非常合理的，以此为基础提出分类器集成方法相比其他方法效果有明显提升。该方法对数据进行特征划分，得到具有差异性的子集，采用适当的核函数和参数对子集分别进行训练和预测，最后将预测结果进行集成，得到最终的预测结果。本研究的创新点有2点：①针对气体图谱数据高维、非线性、样本少的特点，采用支持向量机方法进行分类识别，有别于传统的统计分析方法、神经网络分类器；②不论对哪种输入向量，不同核函数的支持向量机分类结果有所差异，为了融合不同的分类器结果，保证分类的准确性和稳定性，本研究采用了集成的方法。与单一分类器方法来进行糖尿病预测结果相比较，该集成方法的预测性能明显好于单一分类器方法。因而基于特征分类器集成的预测方法为使用医用电子鼻进行糖尿病诊断分析提供了一种有效手段。

参 考 文 献

[1] 宋宝珠,程钊,孙立友,等.中医专家电脑系统研究概况.安徽中医学院学报,1987,14(6):56-62

[2] 郑哲洲,林雪娟.电子鼻在医学诊断中的应用研究.世界科学技术-中医药现代化,2012,14(6):2115-2119

[3] Polat K,G ü ne S.An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease.Digital Signal Processing,2007,17:702-710

[4] Temurtas H,Yumusak N,Temurtas F.A comparative study on diabetes disease diagnosis using neural networks.Expert Systems with Applications,2009,36:8610-8615

[5] Polat K,G ü ne S,Arslan A.A cascade learning system for classification of diabetes disease:Generalized discriminant analysis and least square support vector machine.Expert Systems with Applications,2008,34:482-487

[6] ali ir D,Do antekin E.An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Systems with Applications,2011,38:8311-8315

[7] Thomas G D.Ensemble learning.The handbook of brain theory and neural networks,2002

[8] Rosen B.Ensemble learning using decorrelated neural networks. Connection Science,1996,8:373-384

[9] Wang Ping,Tan Yi,Xie Haibao et al.A novel method for diabetes diagnosis based on electronic nose.Biosensors & Bioelectronics,1997,12(9-10): 1031-1036

[10] 林雪娟,吴青海,李灿东,等.基于气体传感器阵列技术的中医电子鼻研究.世界科学技术-中医药现代化,2012,14(5):1992-1995

[11] 于勇,王俊,周鸣.电子鼻技术的研究进展及其在农产品加工中的应用.浙江大学学报(农业与生命科学版),2003,29(5):579-584

[12] Pal M.Random forest classifier for remote sensing classification. International Journal of Remote Sensing,2005,26:217-222

[13] Breiman L.Random forests.Machine learning,2001,45:5-32

[14] Díaz-Uriarte R,Andres S A.Gene selection and classification of microarray data using random forest.BMC Bioinformatics,2006,7:3

[15] Jiawei Han,Micheline Kamber.数据挖掘概念与技术.范明,孟小峰译.北京:机械工业出版社,2004

[16] 刘玉杰,刘毅慧.基于小波低频系数基因芯片数据的特征提取.生物信息学,2011,3:21

[17] 赖丽娟,吴效明.ICU中发生急性低血压的预测方法研究进展.北京生物医学工程,2010,29(5):538-542

（收稿日期：2015年2月5日）